What's new ?

Sebastien Goasguen

sebgoa@clemson.edu

School of Computing

Clemson University, Clemson, SC

Scientific Associate at CERN

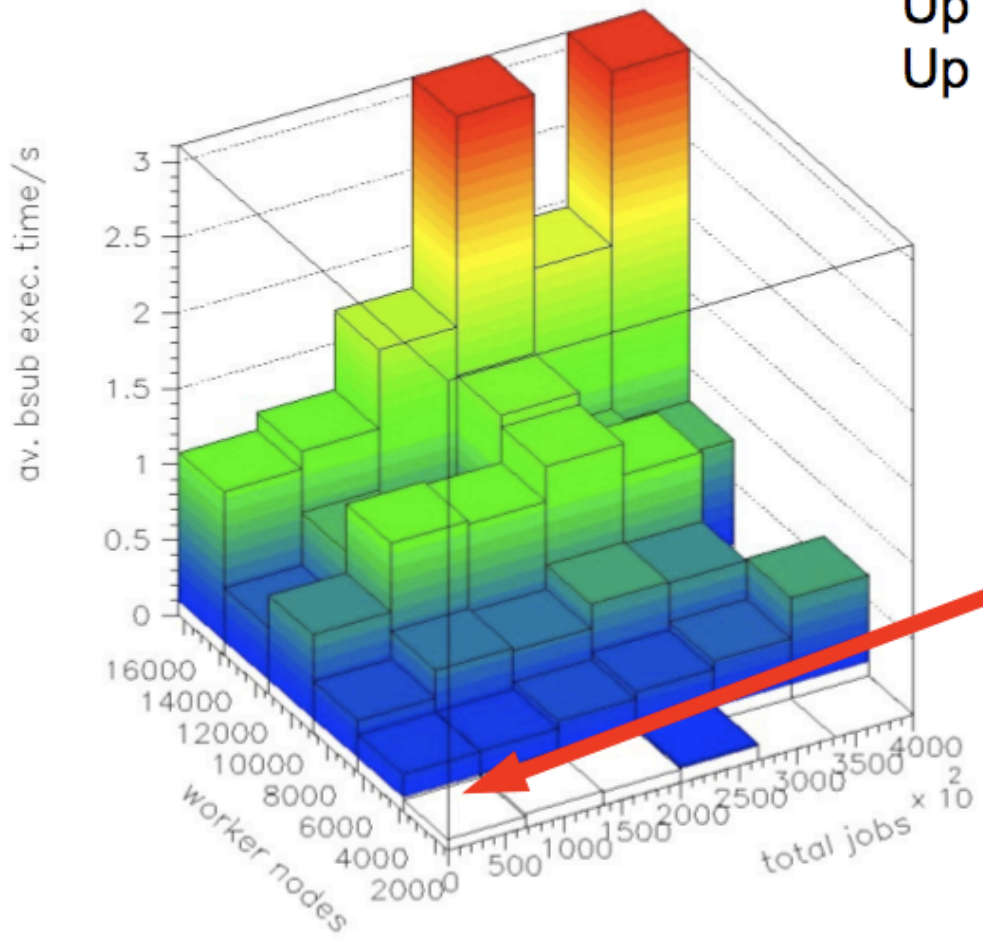
Summer 2009 and Summer 2010

Outline

- **Update on LXCLOUD**
- Recap: Virtual Machine Image Catalogue (VMIC)
- Recap: VMI distribution
- How to start VMs via Engage, NERSC, CERN, Nebraska.
- Forecast for 2011: Openstack

Scalability of the Batch System

Batch system tests: job submission



Up to 15,000 nodes
Up to 400,000 jobs

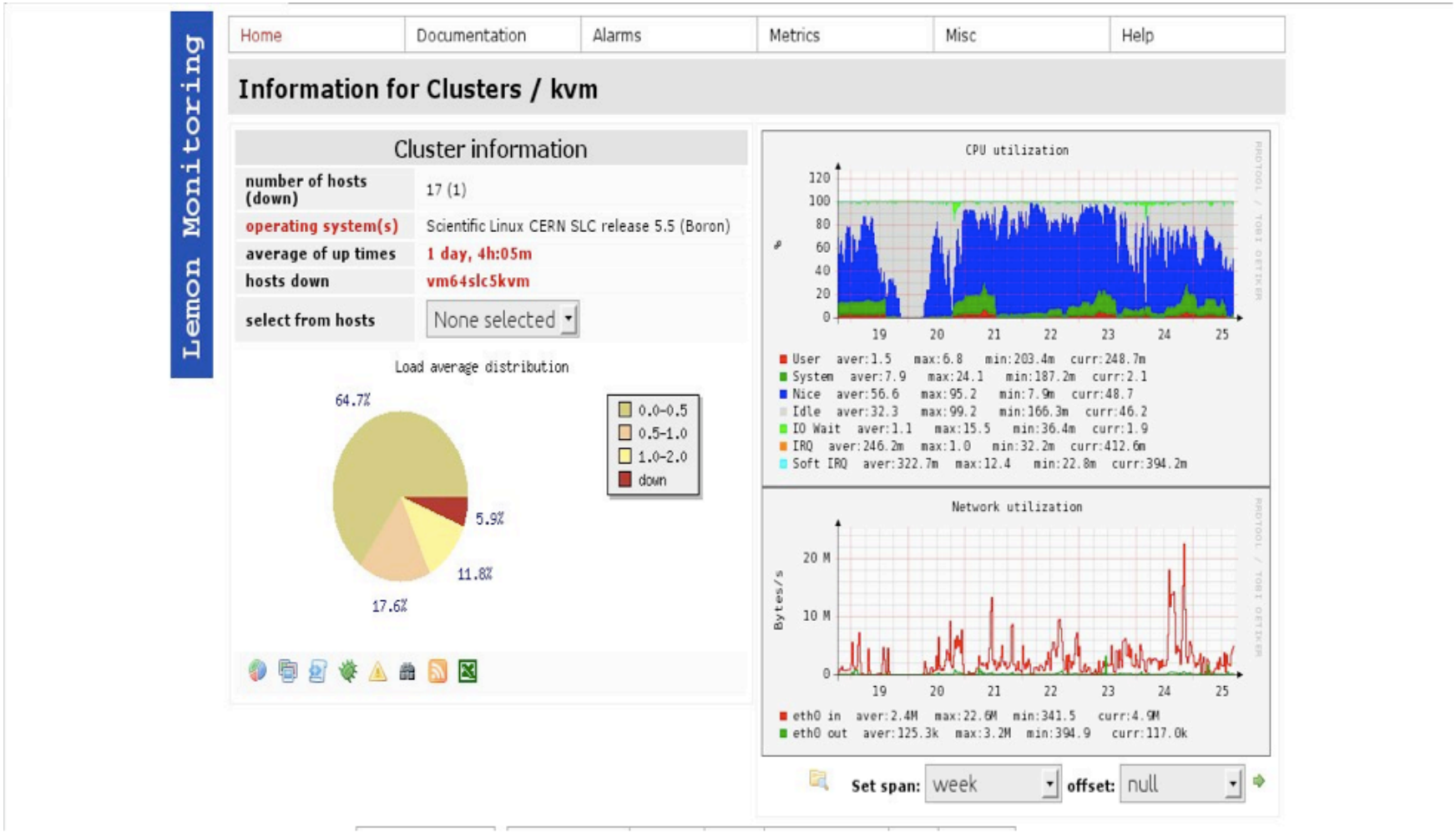
More than **3x** of what is officially supported

Current production system

preliminary

21

Currently in Production



Outline

- Update on LXCLOUD
- **Recap: Virtual Machine Image Catalogue (VMIC)**
- Recap: VMI distribution
- How to start VMs via Engage, NERSC, CERN, Nebraska.
- Forecast for 2011: Openstack

Motivating factors

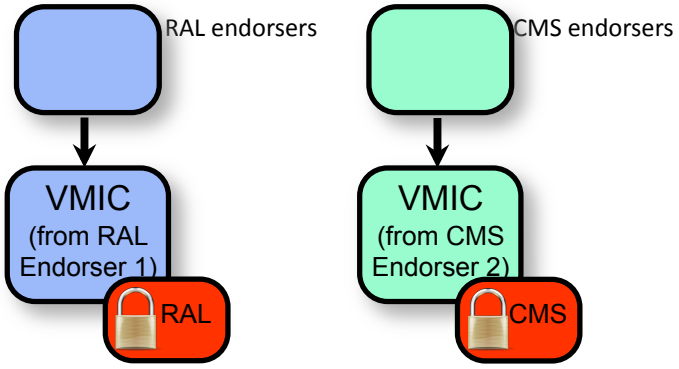
Management of the virtual images (VMIC, HEPiX WG)

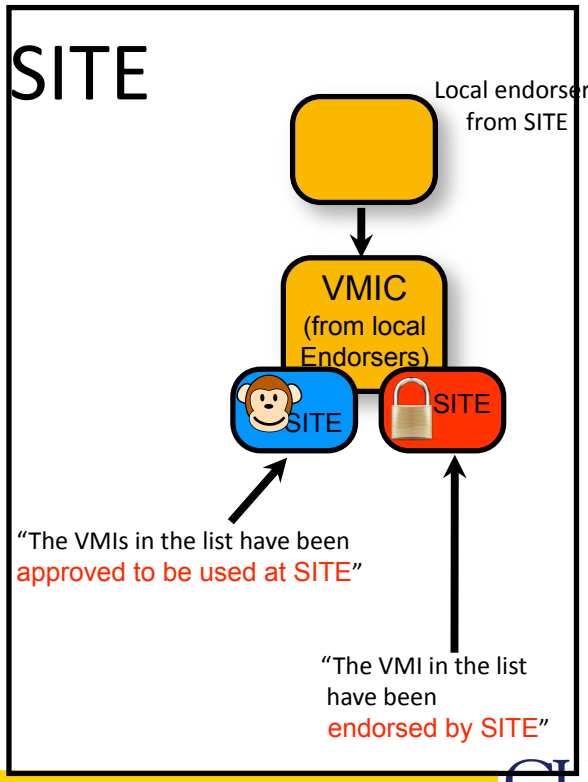
- How can we establish a trusted model to share images?
- What are the requirements on image producers?
- How can we integrate our local image distribution with this?

Transfer of the images to the nodes (LXCLOUD, production)

- How to copy the images to all the hypervisors?
- How to maintain a consistent set of images?
- How to manage changes in the image set?
- How could we optimize network usage?

Endorsed vs approved

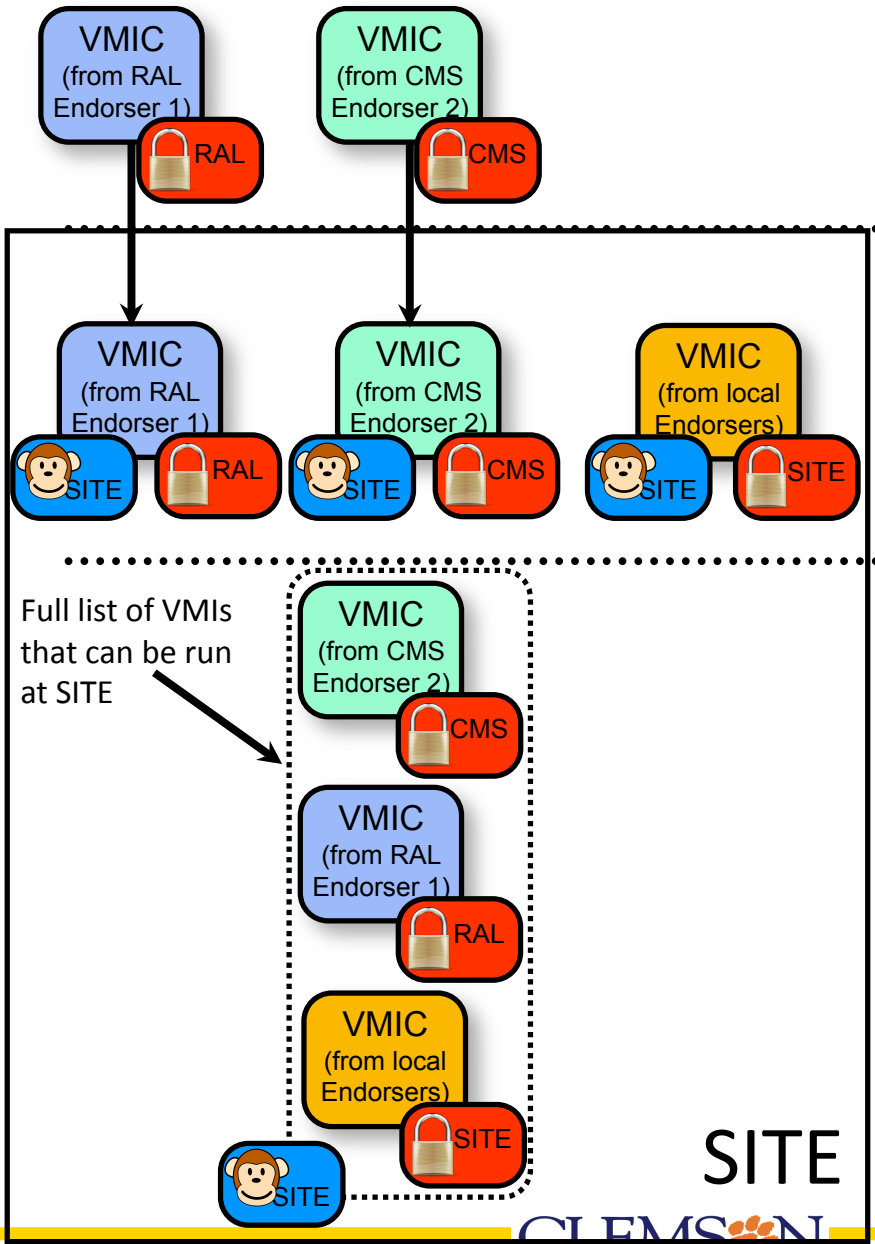




- Endorsed (endorser decision):
  - Role defined in the policy document
  - Scope: VMI production & maintenance
- Approved (site decision):
  - Marks the VMI "valid for use" by the site
  - Scope: operating the VMI

- For a VMI to run, it must be both:
  - Endorsed by an endorser (i.e. Part of the VMIC endorsed)
  - Approved by the local site
- The VMI is run only when the two conditions are met
  - The site has control over VMIs being run
  - The endorser has control over VMIs being produced/endorsed/published

Heterogenous example



1.SITE decides to approve VMIs endorsed by (XYZ) RAL and CMS

2. VMIs are approved(SITE)

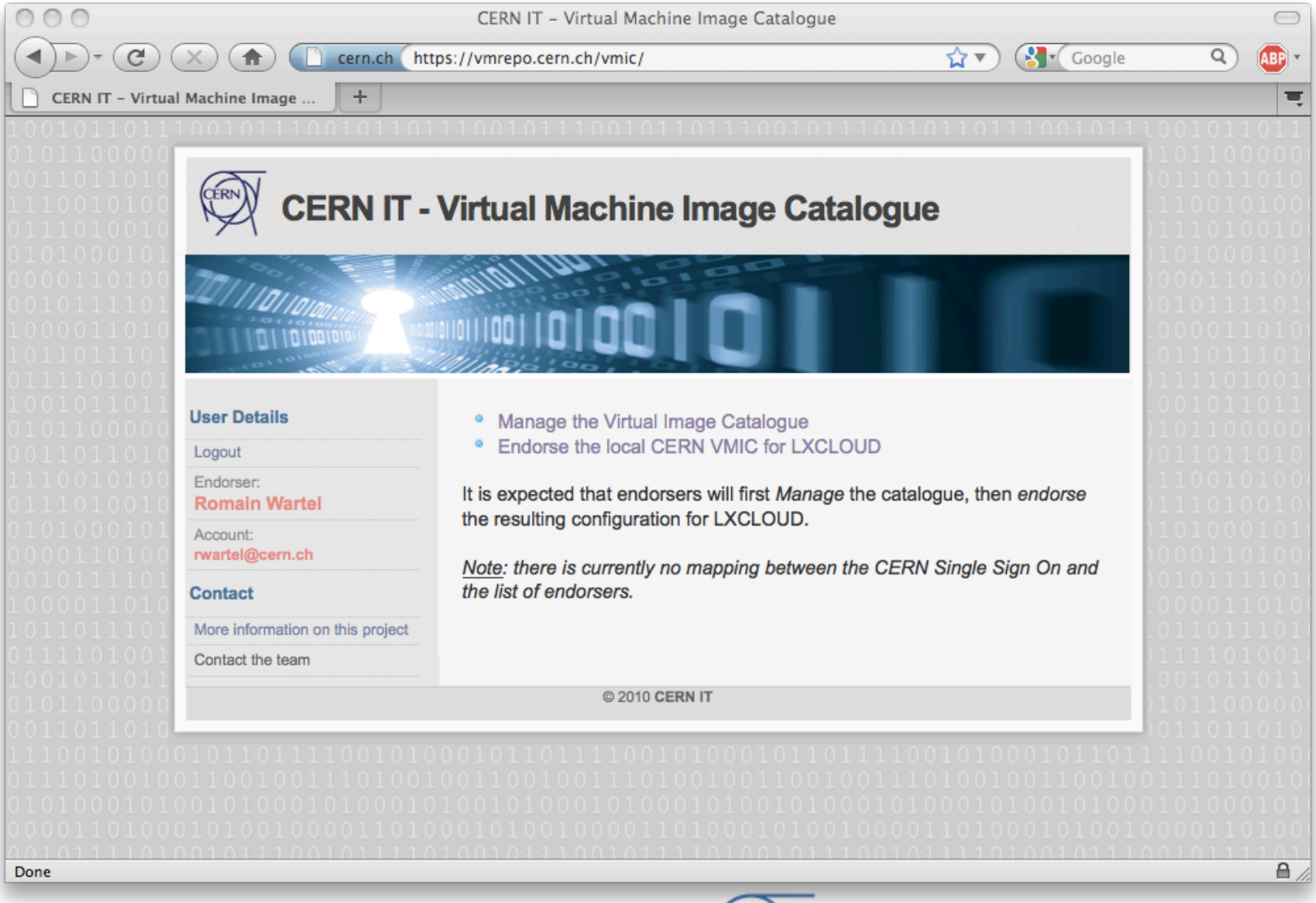
Sites has fine-grained control over VMIs being approved (but can also approve them all)

3. The RAL and CMS VMIs are added to complement the VMIs produced locally

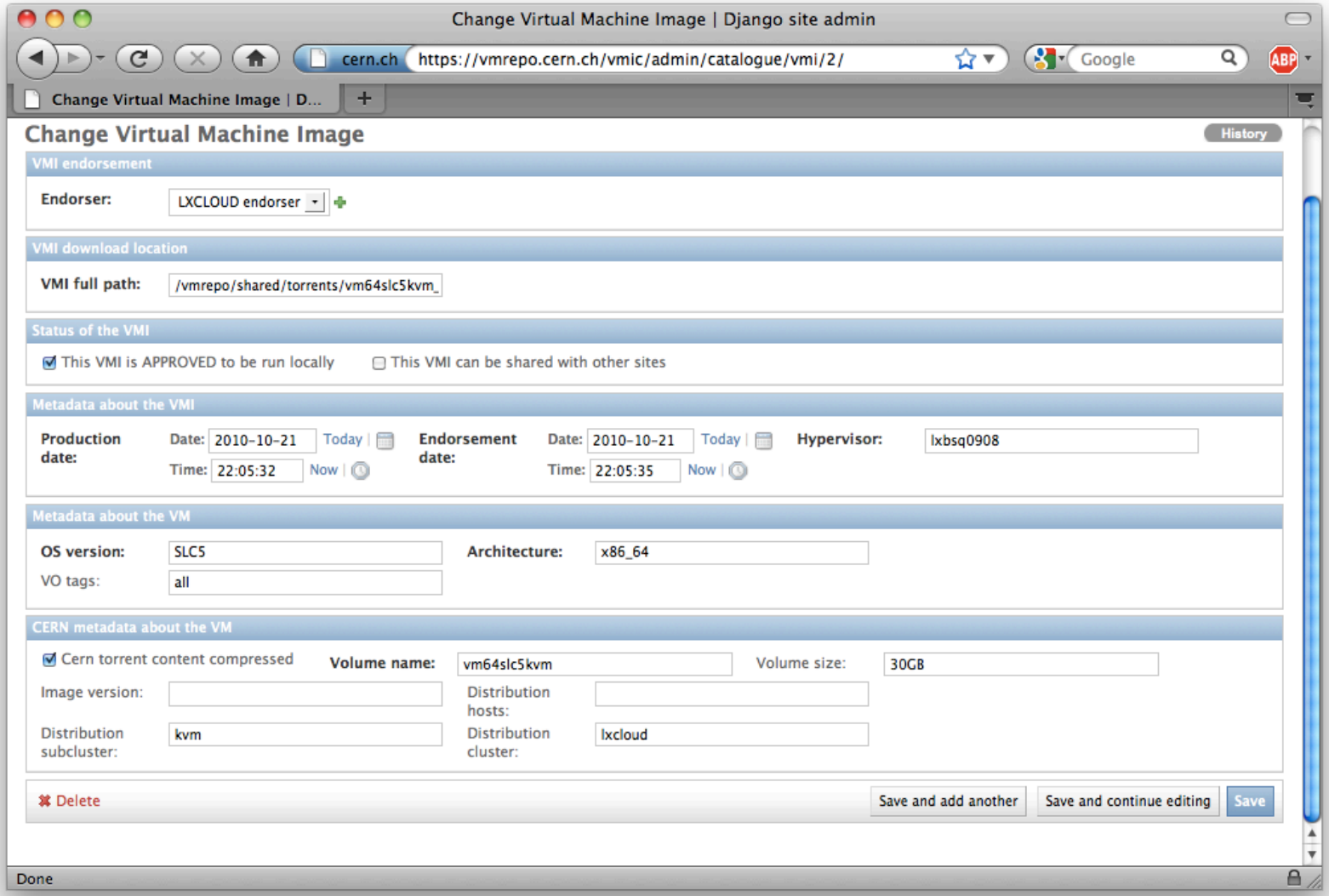
4. The resulting list of VMIs (endorsed by different entities) is approved by the local site

XYZ + SITE + SITE

Implementation



Implementation



VMIC at Clemson

- Current effort
- Take the CERN VMIC code, deploy it a Clemson and tie to local infrastructure
- Potentially support an Engage VMIC

- IMHO, only code that might make sense to put in VDT, if mechanisms are accepted.

Outline

- Update on LXCLOUD
- Recap: Virtual Machine Image Catalogue (VMIC)
- **Recap: VMI distribution**
- How to start VMs via Engage, NERSC, CERN, Nebraska.
- Forecast for 2011: Openstack

Image Distribution

Push:

- Sequential SCP
- logarithmic SCP (scp-wave)
- http://code.google.com/p/scp-wave/
- Upcoming scp tsunami rivals bittorent.

Pull:

- Bittorrent (Romain Wartel, Belmiro Moreira @CERN)

Shared FS :

- NFS, PVFS, Lustre...

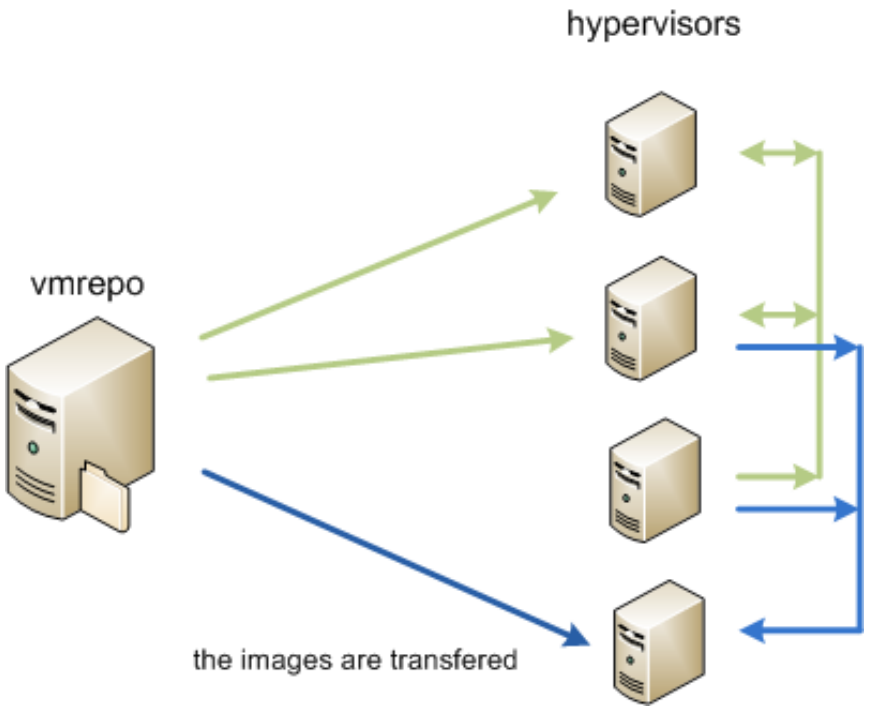


Image distribution over WAN

- What can we use from the VOs to transfer O(10GB) files to multiple sites in a secure way ?
- Where to stage the image on the site ?
  - $OSG_DATA ?
  - SE ?

Bittorrent implementation details

Transfer of the images using Bittorrent

- Central torrent index of trusted images
- Signed list of trusted torrents (in YAML)
- Contains metadata (including hash) of valid torrents

All hypervisors:

- Run a local rtorrent instance
- Download the torrent index on a regular basis, verify its signature
- Select the relevant torrents to be downloaded
- Use the YAML data to download the torrent files, and check their signatures
- Feed the torrent files to rtorrent to download the actual image

Opentracker used as a "booster"

- DHT-only is not sufficient for bootstrapping the P2P network
- DHT and Peer Exchange enabled on all nodes

The torrent list

```
torrent-id_1:
        torrent-name: cernvm_1276695059-img-1277992808.torrent
        torrent-file-sha1: 6b0ca19a41ded48899c17f035760e105049c1118
        download-url: https://vmrepo.cern.ch/torrents/cernvm_1276695059-
img-1277992808.torrent
        torrent-content-hash: 65b6c891528f59e04f6da7621923bba7c4593630
        torrent-content-name: cernvm_1276695059.img.gz
        torrent-content-size: 236.39 MBs
        torrent-creation-time: Thu, 01 Jul 2010 16:00:10 +0200
        torrent-content-compressed: Y
        volume-name: cernvm_1276695059-img
        volume-size: 30
        image-version: 1277992808
        distribution-hosts:
        distribution-subcluster:
        distribution-cluster:
        uuid: cernvm_1276695059-img-1277992808

x509-signature: rsf9Z2bQyXyzNFLOrIi9Jhx96Vfek7gGhXwZXqcTgvKvWP05kzIBT7scV0/
c1y9OokMCWYkn2FdU+ed0euCHhPO8bk4DUNEJ/L6kzHMVPu2Uc5CYkDhyEEPHj0NEIze/
9zb0tBcfuJtD3eC8hHlYaMrP7yjfTzOkr/Wkp3p2zdswk17WjguEtn6ABlsCrR9AwJ2rWMmv76tQwY2MgObz
+kdlpyK26SkmXqv1RnaPQSNPrIj/i4uRl6IJASc0vRofMz3UL4ID7iQ
+rL6aXdzXle4FbjLhTk8CRR8ZIRvLGYLhvnj5ticIJlEyLpnK4rSgfqqGzi8Is3tStDPFipVK2g==
```

Hypervisor Utility: imgdist

```
[root@lxbst0601 scripts]# ./imgdist --list_available
Volume Name                Version          State                Profile
------------------------------------------------------------------------
------------
belmiro.img.gz             1278236546       active               belmiro-
img-1278236546
```

```
[root@lxbst0601 scripts]# ./imgdist --list_archive
Volume Name                Version          State                Profile
------------------------------------------------------------------------
------------
b_1.data                   1278110642       download_ERROR       archive/
b_1-1278110642
b_2.data                   1278167265       download_ERROR       archive/
b_2-1278167265
```

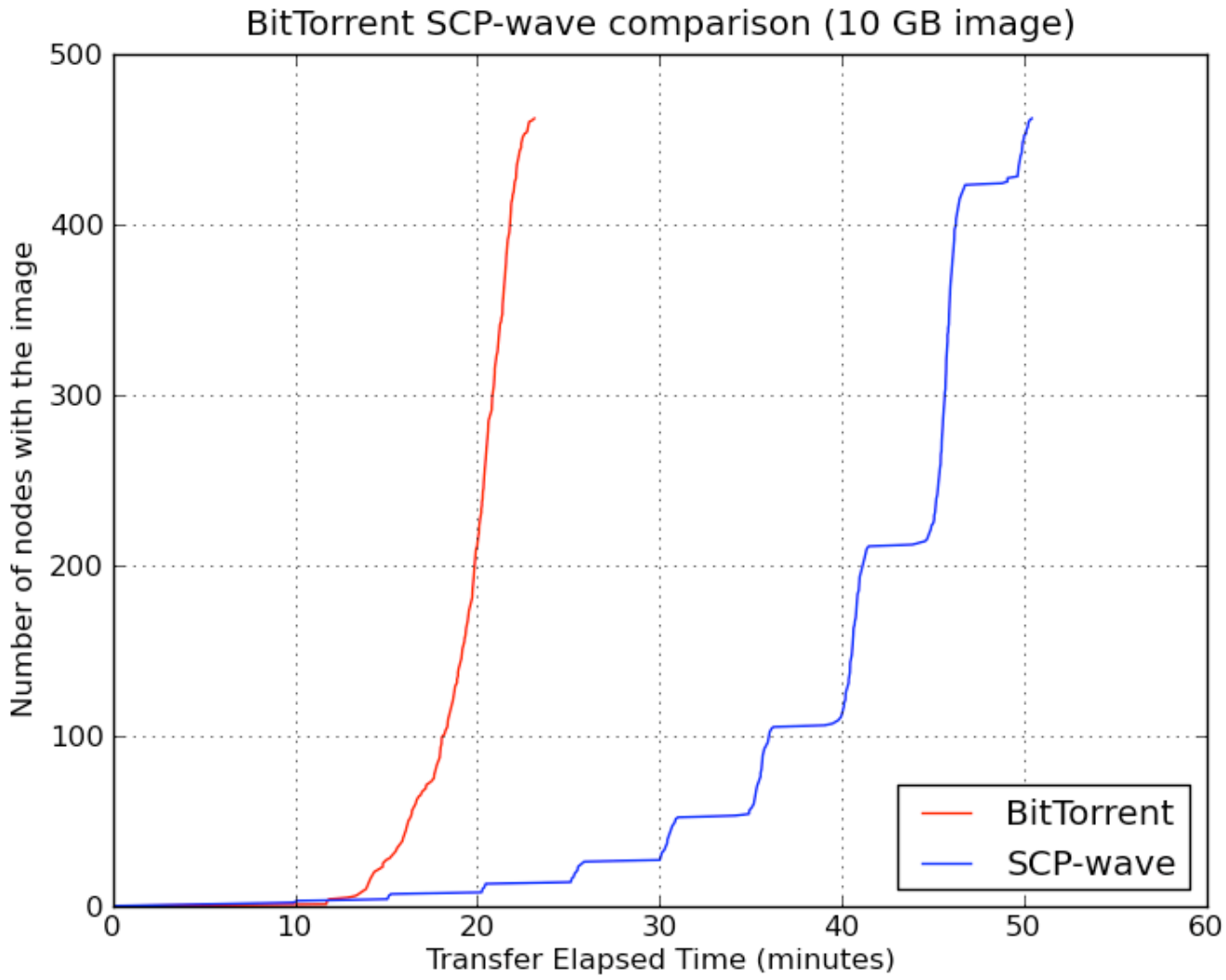
```
[root@lxbst0601 scripts]# ./imgdist --disable
current state: enable
disable
```

```
[root@lxbst0601 scripts]# ./imgdist --list_unknown_files
isfvm64slc5_1276695175-img-1277826366.torrent
vm64slc5_1277964931-img-1277993853.torrent
isfvm64slc5_1276695175.img.gz
```

Current VMIC

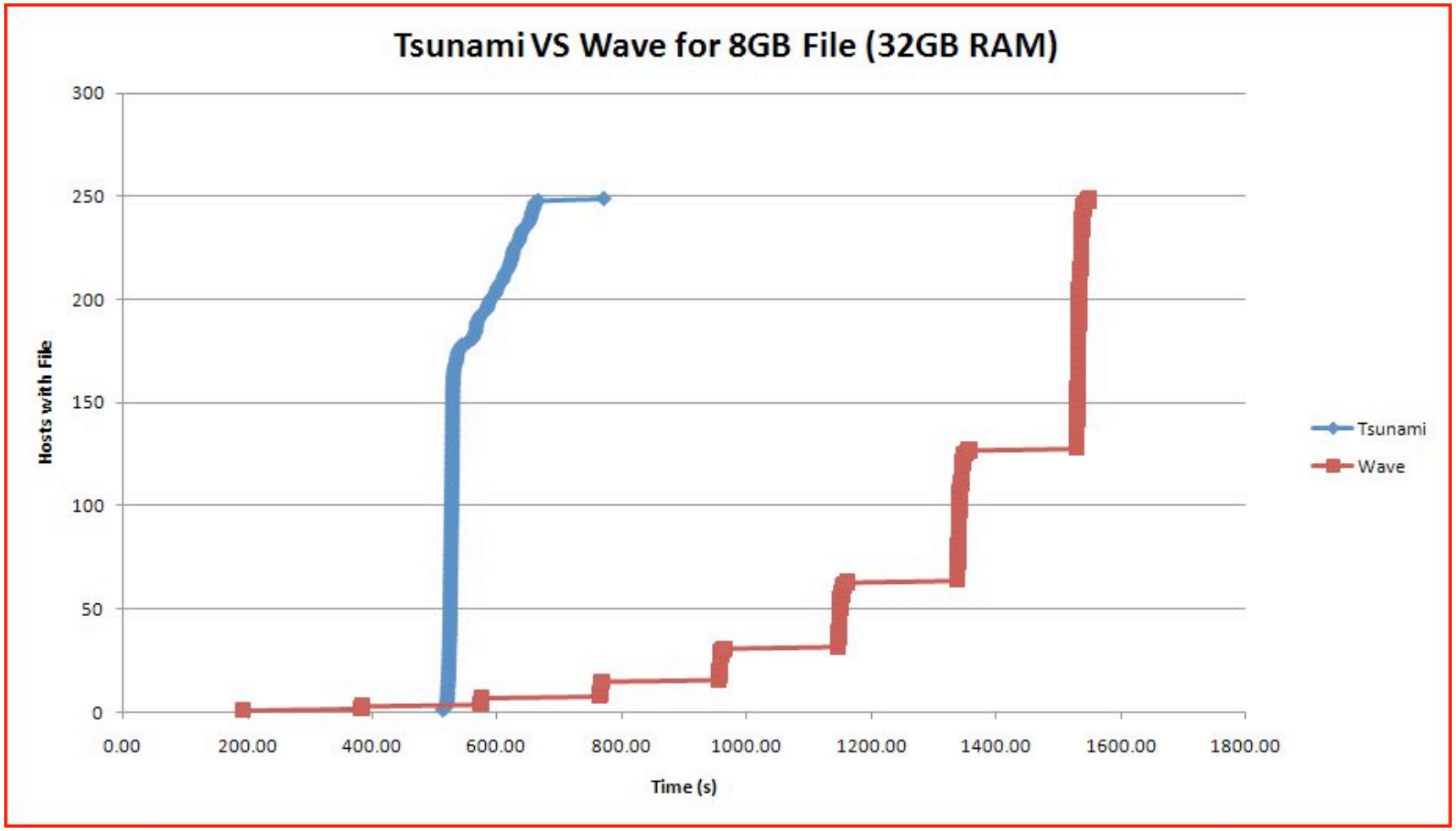
- Only a local one
- Does not contact other **endorser's** VMIC to build a list of **approved** VMs at the site.
- Only connected to the CERN image distribution system, goal of deployment at Clemson is to adapt it to a "smaller" site mechanisms.

Image distribution results



28 Gbit/s aggregate With BT

Released "Tsunami" at: http://code.google.com/code/scp-tsunami



Outline

- Update on LXCLOUD
- Recap: Virtual Machine Image Catalogue (VMIC)
- Recap: VMI distribution
- **How to start VMs via Engage, NERSC, CERN, Nebraska.**
- Forecast for 2011: Openstack

Clemson via Condor

```
[sebgoa@user001 ~]$ more kvmtest.sh
#!/bin/sh
export TMPDIR=/local_scratch
for i in `seq 1 10`
do
kvm -hda "/home/sebgoa/
    microworker2.vm" -m 36 -net nic,vlan=1
    -net user,vlan=1 \
  -nographic -snapshot &
```

```
[sebgoa@user001 ~]$ more condor.kvm
universe = vanilla
notification = never
executable = kvmtest.sh

should_transfer_files = YES
transfer_executable = true
when_to_transfer_output = on_exit

output = out
error = err
log = job.log
```

Clemson via PBS

```
[sebgoa@user001 kvm]$ more
    kvmnew.pbs
#!/bin/bash -l

#PBS -N testkvmnew
#PBS -l nodes=1:ppn=1:intel
#PBS -k oe
#PBS -l walltime=72:00:00

nohup /home/sebgoa/kvm/vmnew.sh
    > vm.out 2>&1 &
sleep 259200
```

```
[sebgoa@user001 kvm]$ more
    vmnew.sh
#!/bin/sh

for line in `cat $PBS_NODEFILE`; do
  echo test
  ssh $line 'nohup /home/sebgoa/
    kvm/start_node3 > startnew.out
    2>&1 &'
  sleep 10
done
```

Engage

```
sebgoa@engage-submit:~/vmtest/inputs$ ls -l
total 265100
-rwxr-xr-x 1 sebgoa external 271187968 2010-11-03 13:32 engage.vm
-rwxr-xr-x 1 sebgoa external       174 2010-11-03 13:27 vmtest.sh
sebgoa@engage-submit:~/vmtest/inputs$ more vmtest.sh
#!/bin/sh
export TMPDIR=/local_scratch

for i in `seq 1 10`
do
kvm -hda "./engage.vm" -m 36 -net nic,vlan=1 -net user,vlan=1 \
   -nographic -snapshot &
done

sleep 172800;
```

Nebraska via condor vm universe…under test

```
universe = grid

grid_resource = condor red.unl.edu red-condor.unl.edu

executable=/bin/date


remote_universe = vanilla

ShouldTransferFiles = YES

WhenToTransferOutput = ON_EXIT

+remote_JobVMType = "KVM"

+remote_JobVMNetworking = false

+remote_VMPARAM_No_Output_VM = true

+remote_JobVM_VCPUS = 1

+remote_JobVMCheckpoint = false

+remote_JobVMMemory = 1536

+remote_VMPARAM_Kvm_Disk = "/var/tmp/OSG_EngageVM/engage.vm:hda:r"

+remote_requirements = ( TARGET.HasVM ) && ( TARGET.VM_Type == "kvm" ) && ( TARGET.VM_AvailNum > 0 )
      && ( VM_Memory

 >= 1536 )
```

NERSC

- Get a regular user account at NERSC via affiliation to a VO (for me STAR).
- Setup the EC2/Eucalyptus environment then use the euca tools to manage instances.
- Not clear how to bundle images…
  - Stuck there, working with STAR to move forward.

CERN

- Opennebula:

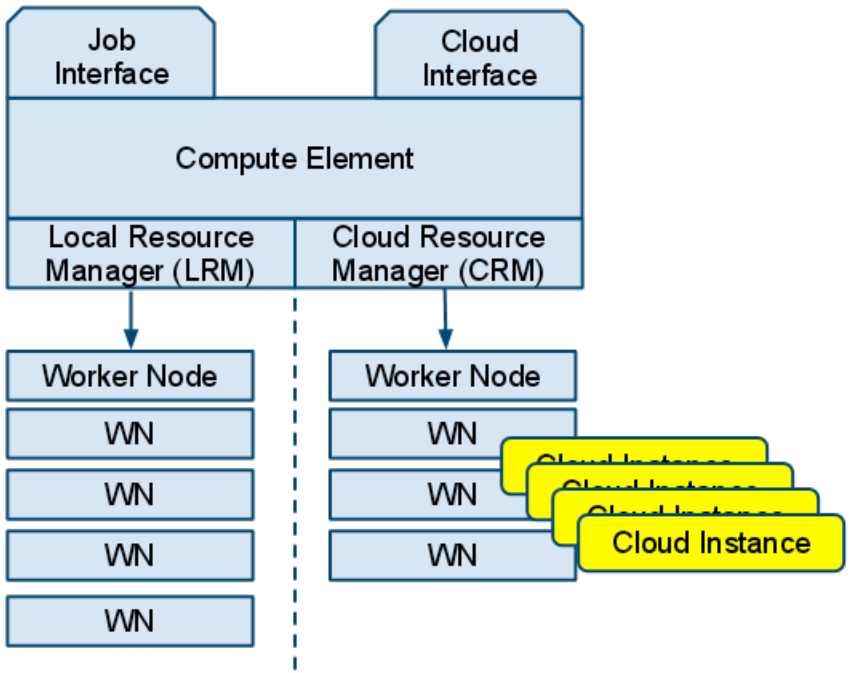
```
onevm create vm.template
onevm delete vm.template
onevm migrate vm.template
```

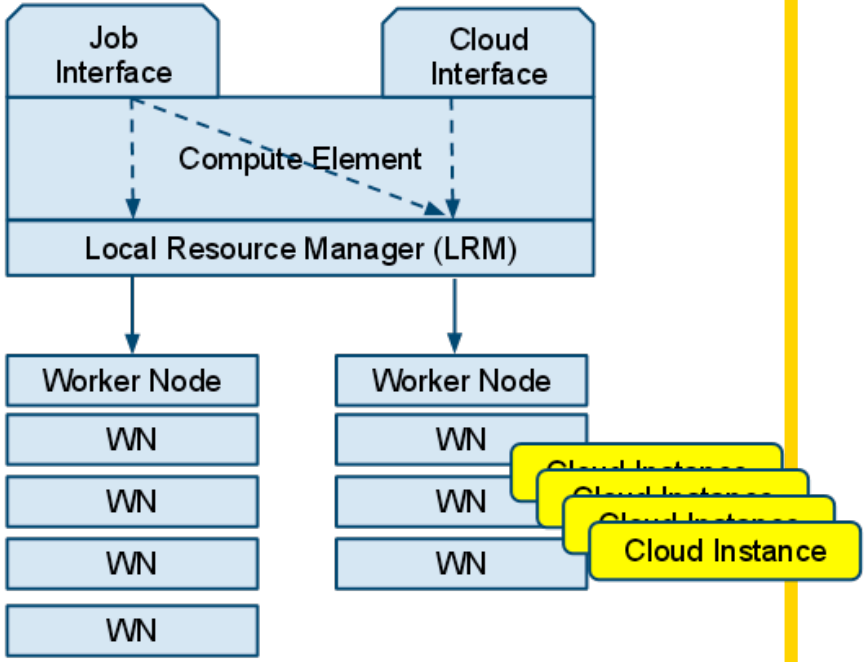
- Can use XML-RPC client, limited cloud API, supposedly can use DEltacloud

What does it mean for OSG

- Can and should treat a VM as a job. Submit a VM to a site using Condor-G is a valid solution no need to deploy cloud API servers.
- IMHO, one does not even need the condor VM universe and can use its default scheduler.
- VM provisioning system become necessary when there is over-provisioning and you want to provide an interactive cloud.

Two configurations for OSG Sites





Outline

- Update on LXCLOUD
- Recap: Virtual Machine Image Catalogue (VMIC)
- Recap: VMI distribution
- How to start VMs via Engage, NERSC, CERN, Nebraska.
- **Forecast for 2011: Openstack**

What is it ?

Cloud computing open source software for compute and storage services.

http://wiki.openstack.org/
http://www.openstack.org/
http://www.computer.org/portal/web/csmediacenter/cloudcom2010/tutorials#

The compute side is from the NASA nebula project, formerly known as novacc.org

The storage side is from rackspace Cloud files service.

Software

- First release (Austin) end of October 2010.
- New release expected first quarter 2011, expected release cycle every 3 months.
- Suggests waiting for Bexar release before testing.
- Compute side all in Python and installation currently highly focused on Ubuntu distro, no rpms.

Openstack Object storage

- Aka SWIFT
- A highly scalable object store, not a file system.

- Architectural overview at:
- http://swift.openstack.org/overview_architecture.html

- Available as a Ubuntu based VM appliance as well as source...

- Good read describing capabilities and limitations at:
- http://adrianotto.com/2010/09/openstack-os-is-great-for/

- noteworthy: not a FS, no quotas, no write to an offset, no append, no file locking, REST API with Python bindings and else.

Openstack Compute

Python code, supports Xen,KVM, UML, Qemu and dev about Hyper-V.

Inside story is that nova gave code to Eucalyptus project but when euca created start-up, they refused the NASA patch and NASA and decided to give code to Openstack in open source. That said a few things look like eucalyptus and you can use the euca tools to manage VMs.

Preferred installation on Ubuntu, Live CD also availbable (note: could be quick easy way to test).

Openstack compute dependencies

```
yum -y install dnsmasq vblade kpartx kvm gawk iptables
ebtables bzr screen euca2ools curl rabbitmq-server gcc
gcc-c++ autoconf automake swig openldap openldap-servers
nginx python26 python26-devel python26-distribute git
openssl-devel python26-tools mysql-server qemu kmod-kvm
libxml2 libxslt libxslt-devel mysql-develeasy_install-2.6
twisted sqlalchemy mox greenlet carrot python-daemon
eventlet tornado IPy routes lxml MySQL-python sphinx boto
webob easy_install-2.6 python-daemon==1.5.5
```

rabbitmq-server is the core publish subscribe system that handles every event synchronously. rabbitmq is an implementation of AMQP. carrot is the python binding for rabbitmq. vblade is the ATA over ethernet which enables you to create a cheap SAN.

Suspect lots of changes in Bexbar release.

Openstack compute

• A client publishes requests for VM instances, hypervisors listens for requests and grabs them from the queue in a FIFO way. (note: Not sure if other type of scheduling is possible)

•More decoupled than opennebula and ISF and theoritically more scalable (note: yet to be proven in practice, talk at CERN by NASA indicated only 100 VMs tests)

•Networking seems more involved/flexible:
•http://nova.openstack.org/adminguide/managing.networks.html

•User management and instance management akin to EC2/ Eucalyptus. Image Files created with separate kernel and ramdisk, not using a single raw image file.

Conclusions

- Things keep on moving
- VMIC being deployed at Clemson
- Various mechanisms tested to start VM at multiple OSG sites.
- Openstack may overtake everything
- Lessons:
  - Highly asynchronous systems preferred
  - Treat data / events as streams and adapt
  - Don't be too deterministic